Protein Science (2000), *9*:1382–1390. Cambridge University Press. Printed in the USA.

FOR THE RECORD

HYR, an extracellular module involved in cellular adhesion and related to the immunoglobulin-like fold

ISABELLE CALLEBAUT,[1] DELPHINE GILGÈS,[2] ISABELLE VIGON,[2] AND JEAN-PAUL MORNON[1]

[1]Systèmes Moléculaires & Biologie Structurale, LMCP, CNRS UMR 7590, Universités Paris 6 et Paris 7, Paris, France
[2]INSERM U474, Hôpital Henri Mondor, Créteil, France

(RECEIVED January 25, 2000; FINAL REVISION April 11, 2000; ACCEPTED May 5, 2000)

Abstract: Domains belonging to the immunoglobulin-like fold are responsible for a wide variety of molecular recognition processes. Here we describe a new family of domains, the HYR family, which is predicted to belong to this fold, and which appears to be involved in cellular adhesion. HYR domains were identified in several eukaryotic proteins, often associated with Complement Control Protein (CCP) modules or arranged in multiple copies. Our analysis provides a sequence and structural basis for understanding the role of these domains in interaction mechanisms and leads to further characterization of heretofore undescribed repeated domains with similar folds found in several bacterial proteins involved in enzymatic activities (some chitinases) or in cell surface adhesion (streptococcal C-alpha antigen).

Keywords: adhesion; complement control protein; fibronectin type III; hyalin; hydrophobic cluster analysis; iterative database search; polycystic kidney disease

Eukaryotic extracellular proteins are mostly composed of multiple modules, each of which can generally be found in a wide variety of proteins with different functions (Bork et al., 1996). One of the more frequently occurring of these modules is the fibronectin type III (Fn3) domain, which is found in many extracellular proteins, in the extracellular parts of several membrane–receptor proteins, and also in some intracellular proteins such as the muscle-associated titin (Bork & Doolittle, 1992). Fn3 domains have also been found in bacteria but are largely limited to carbohydrate-splitting enzymes (Bork & Doolittle, 1992; Little et al., 1994). Fn3 domains form a distinct superfamily within the immunoglobulin-like fold. Their structure consists of a seven-stranded β-sandwich, with two sheets A-B-E and C′-C-F-G packed face to face (Fig. 1) (Baron et al., 1992; Leahy et al., 1992; Bork et al., 1994; Halaby et al., 1999). Another domain with a Greek key β-sandwich topology very similar to that observed for Fn3 domains is the polycystic kidney disease (PKD) domain, originally found in 16 copies in polycystin-1, a protein encoded by the PKD1 gene, which is mutated in autosomal dominant polycystic kidney disease (ADPKD) (Bycroft et al., 1999; Fig. 1). PKD domains are present in a num-

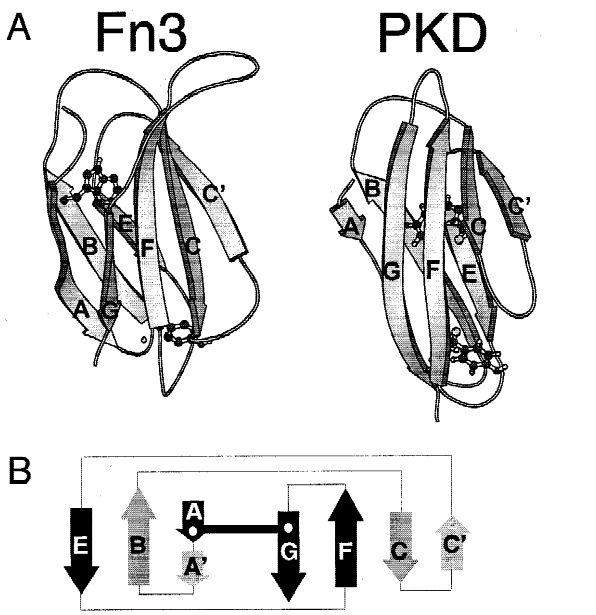

Fig. 1. A: Molscript representation (Kraulis, 1991) of the three-dimensional structures of the tenth Fn3 domain of human fibronectin (Protein Data Bank (PDB) 1FNF) and of a PKD domain of human polycystin-1 (PDB 1B4R), showing the secondary structure elements. β-Strands are labeled A to G and are organized in two sheets of three (ABE) and four (GFCC′) antiparallel β strands. The conserved tyrosine at the beginning of strand βF is shown, as well as the two highly conserved tryptophane residues in strands βB (Fn3) and βC (PKD), respectively. **B:** Schematic diagram of the predicted fold of the HYR domain. β-Strands shaded gray indicate the different labeling possibilities relative to the predicted positions (see text and Fig. 3).

Reprint requests to: Isabelle Callebaut, Systèmes Moléculaires & Biologie Structurale, LMCP, CNRS UMR 7590, Universités Paris 6 et Paris 7, Case 115, 4 place Jussieu 75252 Paris Cedex 05, France; e-mail: callebau@lmcp.jussieu.fr.

ber of eukaryotic proteins and are also found in bacterial collagenases and proteases as well as in a surface-layer protein from *Methanothermus*. Interestingly, PKD domains are often observed, within a same protein sequence, in association with Fn3 domains.

Together with other superfamilies within the immunoglobulin-like fold, such as the immunoglobulin-like (IgSF) and cadherin (CA) superfamilies, domains from the Fn3 superfamily are widely implicated in specific molecular recognition processes and participate in cell adhesion through cell/cell interactions as well as interactions between cells and the extracellular matrix (Jones, 1996). Ligand binding sites are formed by surface loops as well as by parts of the β-sheets. A well-known example of a loop involved in integrin binding is the "RGD" loop, linking strand βF to strand βG (loop F–G) of the 10th fibronectin Fn3 module (Bork et al., 1994; Campbell & Spitzfaden, 1994; Harpaz & Chotia, 1994). On another hand, the CFG face of the immunoglobulin-like β-barrel appears to play a major role in the adhesive interaction, irrespective of the nature of the ligand [e.g., in VCAM-1 for its interaction with the integrin VLA-4 (Jones et al., 1995), in CD2 for its interaction with CD58/CD48 (Bodian et al., 1994), in cadherin for homophilic adhesion (Overduin et al., 1995; Shapiro et al., 1995a)]. Finally, it is worth noting that two consecutive domains are often involved in the binding function, as illustrated by neuroglian Fn3 domains, which display an ion-binding cleft between them (Huber et al., 1994).

Here, we describe a new family of extracellular protein modules likely to play an important role in cellular adhesion, as these modules are responsible for the interaction of hyalin, a protein of the echinoderm extra-embryonic matrix, with its cell surface receptor (Wessel et al., 1998). As hyalin is composed exclusively of repeats of this domain, we will refer to this heretofore undescribed domain as the HYR module (hyalin repeat). Moreover, we show that this domain family probably corresponds to a new superfamily within the immunoglobulin-like fold, as suggested on the one hand by the sequence similarities it shares with members of the Fn3 and PKD families, and on the other hand by a distinct sequence pattern conservation relative to these families. Interestingly, we additionally point out the presence in some bacterial proteins of repeated domains, which are also predicted to belong to the immunoglobulin-like fold, thereby linking HYR modules to Fn3 domains.

Results and discussion: *Identification of the HYR family:* We started our analysis from the sequence of a recently cloned murine cDNA encoding a large protein (3,567 amino acids) called Polydom, which contains a Von Willebrand factor type A (VWA) domain and a pentraxin (PTX) domain, as well as multiple copies of complement control protein (CCP) and Epidermal Growth Factor (EGF)-like modules (Fig. 2; Gilgès et al., unpubl. obs.). Using hydrophobic cluster analysis (Callebaut et al., 1997), a duplicated domain of ~80 amino acids corresponding to the HYR module described herein was first identified between the third and fourth CCP modules of Polydom (32% identity, BLAST 2 sequences E-value of 8×10^{-5}). Then, using both domains as queries in PSI-BLAST searches (Altschul et al., 1997), the HYR module was significantly detected after convergence in several proteins (Figs. 2, 3; see the details of the PSI-BLAST results in the Fig. 3 caption).

In the hyalin protein, which is the major constituent of the hyalin layer of echinoderm embryos, the HYR module corresponds to a repeated domain, which constitutes the entirety of the protein (Wessel et al., 1998; Fig. 2). The hyalin layer is an extra-embryonic

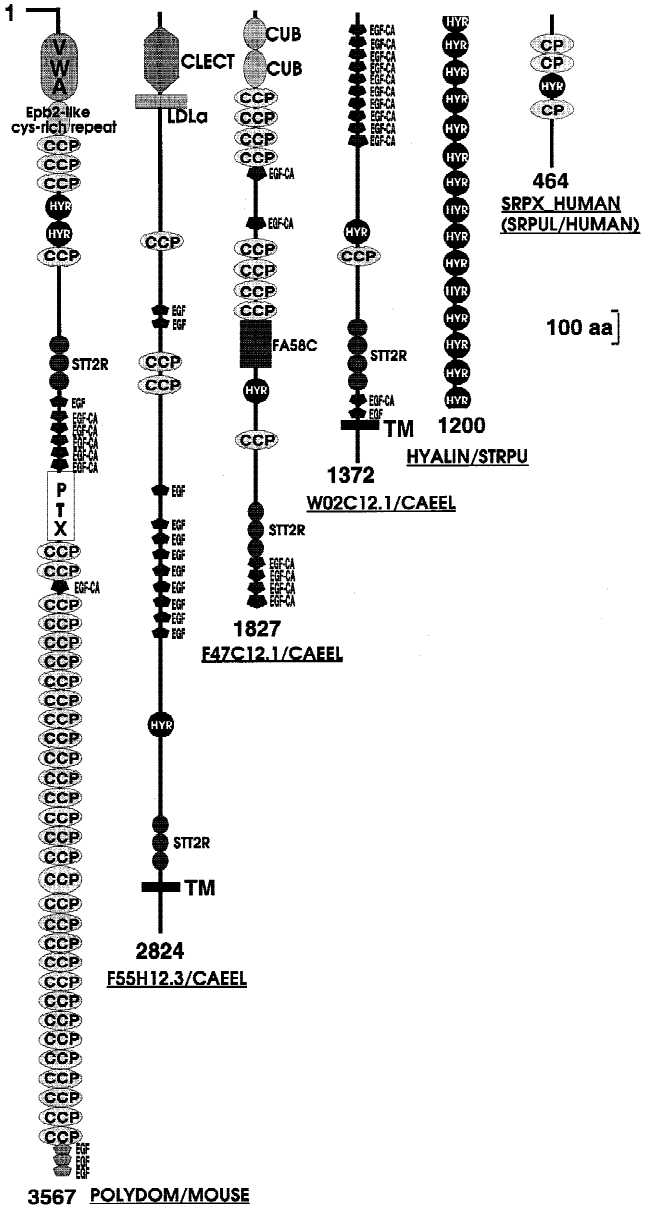

Fig. 2. Modular architecture of proteins containing the HYR module. Genbank and SwissProt accession numbers are given in Figure 3. Other domains were identified against the Pfam (Bateman et al., 1999), Prosite (Hofmann et al., 1999), and SMART (Ponting et al., 1999) databases and are abbreviated as follows (Bork & Bairoch, 1995): CCP, Complement Control Protein; CUB, C1r/C1s, uEGF, bone morphogenetic protein; FA58C, F5/8 type C, also known as discoidin domain (DS); EGF, Epidermal Growth Factor-like; EGF-CA, calcium-binding Epidermal Growth Factor-like; HYR, Hyalin Repeat; LDLa, Low-Density Lipoprotein receptor domain class A; PTX, Pentraxin; VWA, Von Willebrand factor type A. Putative transmembrane regions (TM) are indicated. An additional Cys-rich domain (STT2R, Similar To Thyroglobulin-like 2 Repeats), composed of three repeats of a two cysteine basic module sharing similarities with thyroglobulin-like 2 repeat is found in three or four copies in the three *C. elegans* hypothetical proteins as well as in the Polydom protein. One of the CCP modules of mouse Polydom is larger due to an insertion of considerable length within this domain. Note also that the HYR sequence repeat of hyalin is "switched" relative to the structural repeat (see also Fig. 3 and text).

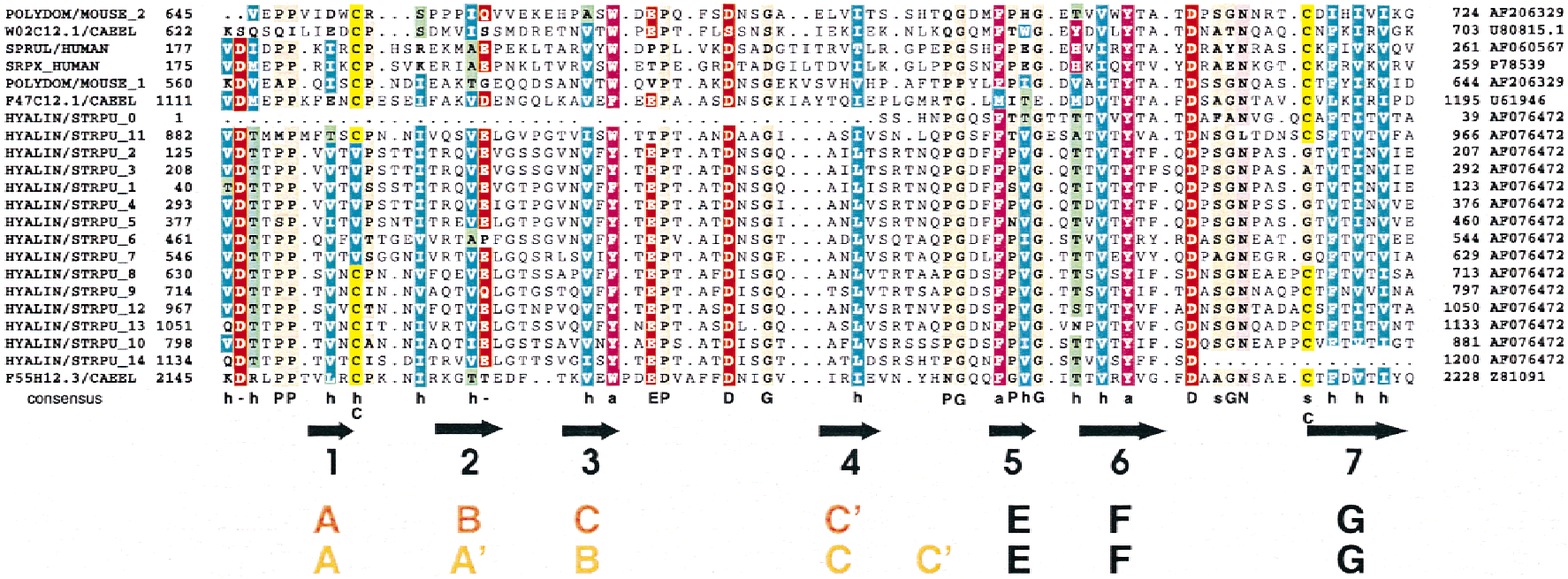

Fig. 3. Multiple alignment of HYR domain sequences. PSI-BLAST searches (BLAST 2.0.10; E-value threshold 0.001) of NCBI nonredundant database (nr; 436 362 sequences) using the Polydom first and second HYR domains (aa 560 to 644 and aa 645 to 724) as queries revealed similarities with the following proteins after convergence by iterations 2 and 4, respectively: SRPX [from 4×10^{-29} to 8×10^{-26} (first HYR domain) and 5×10^{-24} to 6×10^{-23} (second HYR domain)], *Strongylocentrotus purpuratus* hyalin (14 hits ranging from 8×10^{-20} to 6×10^{-10} (first HYR domain) and from 1×10^{-17} to 2×10^{-7} (second HYR domain) and several hypothetical proteins from *C. elegans*: F47C12.1 (4×10^{-22}, first HYR domain; 2×10^{-18}, second HYR domain), W02C12.1 (5×10^{-16}, first HYR domain; 9×10^{-22}, second HYR domain), and F55H12.3 (2×10^{-18}, first HYR domain; 5×10^{-18}, second HYR domain). The repeats of hyalin from *Lytechinus variegatus* (AF076250) are not indicated for clarity. Swiss-Prot and GenBank accession numbers are given at the ends of the sequences. Secondary structure predictions using the JPred server (Cuff et al., 1998) are given beneath the alignment (arrow: extended or β-strand). A consensus line is also shown: h, hydrophobic residues (V, I, L, M, F, Y, W; in green, and which can sometimes be substituted by S, C, T or A); a, aromatic residues (Y, F, W; in purple); –, negatively charged residues (D, E; in red); s, small residues (A, S, G; in beige); cysteine residues (C; in yellow). Note that two positions at the beginning and end of the HYR domain are often occupied simultaneously by cysteine residues, suggesting that these two residues probably form a disulfide bond. This hypothesis is supported by the overall architecture of the predicted immunoglobulin-like fold. STRPU, *S. purpuratus*; CAEEL, *C. elegans*. Figures 3 and 4 were prepared using ESPript (Gouet et al., 1999).

matrix that acts as a substrate for cell adhesion throughout early development. As hyalin is composed exclusively of HYR modules, and as these have been shown to contain the ligand for the hyalin cell surface receptor (Wessel et al., 1998), the HYR module can also be expected to play a direct role in cellular adhesion in other proteins in which it is present. It should be noted that calcium is involved in the aggregation of hyalin monomers in high molecular weight core particles (Wessel et al., 1998). This divalent cation might therefore also play a role in the structure/function of HYR modules.

The HYR module also appears to be frequently associated with CCP modules (Fig. 2). These modules, also known as Short Consensus Repeats (SCR) or sushi domains, are mainly found in various complement regulatory proteins known to interact with components C3b and/or C4b (Reid & Day, 1989). For instance, CCP modules flank the HYR module of SRPX (sushi-repeat-containing protein, X chromosome), a protein encoded by a gene that is deleted in patients with X-linked retinitis pigmentosa, and which is thought to be located at the photoreceptor cell surface (Meindl et al., 1995).

As shown in Figure 3, HYR domains contain highly conserved residues. Three conserved aromatic amino acids should participate in the packing of the hydrophobic core. Three acidic residues are also highly conserved, suggesting that they may play a specific role, possibly in a cation-binding function. Two positions at the beginning and end of the HYR domain are often simultaneously occupied by cysteine residues, suggesting that these two residues probably form a disulfide bond. Consistent with this hypothesis, when cysteines are absent, these two positions are occupied by a valine (first position) and an alanine or a glycine (second position), all of which are small, uncharged residues that can fill buried positions. Secondary structure predictions on the multiple alignment using the JPred server (Cuff et al., 1998) indicate an all-β fold including seven β-strands (Fig. 3).

Relationship to the immunoglobulin-like superfold: Due to its position (flanked by CCP modules; Fig. 2) and to its length, the HYR domain of SRPX was first suggested to be a divergent CCP module (Meindl et al., 1995). Although both modules should have a similar secondary structure pattern [seven β-strands, as predicted for HYR (Fig. 3), or deduced for CCP from an experimental structure (Barlow et al., 1993)], the analysis presented here clearly distinguishes HYR from CCP modules for several reasons. First, the HYR module contains only two cysteines located N- and C-terminal to the module and not absolutely conserved within the family (Fig. 3), while the CCP modules possess four cysteines that form an intramodule disulfide bridge in a 1-3, 2-4 pattern (Barlow et al., 1993). Second, CCP modules were not detected using sensitive profile-like procedures such as PSI-BLAST (Altschul et al., 1997) or HMMER (Eddy, 1998) with the sequences depicted in Figure 3, even considering hits below the significance level. Finally, the key residues of the two modules, including hydrophobic amino acids, which probably participate in their compact cores, are clearly different.

Using PSI-BLAST, similar searches were performed using the tandem of Polydom HYR domains instead of each domain considered separately. Interestingly, in addition to the above-mentioned members of the HYR family (reported in Figs. 2, 3), significant matches were also observed with other domains, some of which correspond to well-established Fn3, PKD or CA domains (Fig. 4; Table 1). The observed similarities match highly conserved motifs located at the end of the HYR modules, including the two last predicted β-strands (strands β6 and β7) and the loop connecting them (Fig. 4). Moreover, they correspond in the matching proteins to repeated domains with a length similar to the HYR domain (~80–100 amino acids) and which, in some cases, constitute most or even the all of the protein, as in hyalin (Table 1). In these cases, as in hyalin, the sequence repeat is frequently switched relative to the structural repeat, suggesting that the N-terminal end might participate in a unique globular structural unit with the C-terminal end via a secondary structure exchange mechanism (like that observed in chaperon-adhesin complexes; Choudhury et al., 1999; Sauer et al., 1999) occurring in a circular arrangement or in oligomers.

As observed in Figure 4, the residue pattern of strand β6 is highly conserved in HYR domains relative to Fn3 domains, and more particularly PKD domains (strand βF). This pattern includes a glycine at the beginning of the strand, followed by an alternation of nonhydrophobic/hydrophobic amino acids typical of extended structures, and ended by a conserved acidic residue. This pattern is immediately followed by a highly conserved tripeptide (consensus sequence [AS]G[NQE]), which is located in the FG loop, and which matches, in the fibronectin 10th module, the integrin-binding tripeptide RGD. The conservation of strand βF of HYR domains relative to those of the Fn3 and PKD superfamilies is consistent with the fact that it is the only strand to retain clearly conserved hydrophobic features in all the immunoglobulin-like folds (Bork et al., 1994; Halaby et al., 1999). However, in HYR domains, the conserved aromatic residue does not occupy the first hydrophobic position (as in Fn3 or PKD domains) but the third one. It should be noted that intermediate states can be observed in some repeated domains sharing similarities with HYR domains, in which these two hydrophobic positions (first and third) are both occupied by aromatic residues (Fig. 4).

These results and the predicted secondary structure patterns (7 β-strands, Fig. 3), therefore, clearly indicate that the structure of HYR modules as well as repeated modules found in several bacterial proteins described in Table 1 and Figure 4 should, in fact, correspond to immunoglobulin-like folds that share sequence similarities (limited to the C-terminal regions) with three different superfamilies adopting the same immunoglobulin-like superfold (namely, the Fn3, PKD, and CA superfamilies). Moreover, Fn3, PKD, and HYR domains share a specific amino acid composition rich in light amino acids (A, G, S, and T comprise approximately a third of the residues of HYR domains) and unusual for all β-proteins (Bycroft et al., 1999; also see footnote to Table 1).

Although the last two strands can unambiguously be linked to strands βF and βG on the basis of sequence similarities with Fn3 and PKD domains (Fig. 4), HYR modules nonetheless have very different sequence patterns in the N-terminal parts relative to Fn3 and PKD domains, just as these latter two domains also largely differ in the sequence characteristics of their N-terminal 5 β-strands. The different N-terminal conserved sequence patterns of HYR, Fn3, and PKD domains also distinguish from those of other repeated domains described in Table 1. Consequently, the HYR family of domains appears to be a distinct superfamily within the immunoglobulin-like fold. The clear conservation of structural elements participating to the CFG face relative to cell adhesion molecules makes it likely that this face could also play an important role in the HYR adhesive function. The most probable strand assignment for HYR domains (1/A, 2/A′, 3/B, 4/C, 5/E; Figs. 1, 3)

Table 1. *Summary of the PSI-BLAST results using the two HYR domains of mouse Polydom as query, after three iterations [BLASTP 2.0.6, nr database (425,089 sequences)]*[a]

Abbreviation	Name	Species	Acc number (aa number)	Matching domains	Number	Position(s)	E-value	Other identified domains
CHIB/CLOPA	Chitinase B	*Clostridium paraputrificum*	dbj|BAA23796 (831 aa)	nd	2	622–699/700–783	2×10^{-31}	GH18
CHIA/CLOPA	Chitinase A	*Clostridium paraputrificum*	dbj|BAA34922 (832 aa)	nd	2	624–701/702–785	1×10^{-29}	GH18
YEST/BACHA	YesT	*Bacillus halodurans*	dbj|BAA75373.1 (688 aa)	nd	2	444–528/529–607	5×10^{-26}	Fn3
YO13/BPL2	Hypothetical protein	*Acholeplasma* phage L2	sp|P42548 (738 aa)	nd	3	373–455/456–538/539–622	7×10^{-25}	
CHIA/VIBHA	Chitinase A	*Vibrio harveyi*	gi|1763985 (729 aa)	nd	2	352–429/430–506	2×10^{-22}	GH18
SC4C6.20C/STRCO	Putative glycosyl hydrolase	*Streptomyces coelicolor*	emb|CAB45584.1 (728 aa)	nd	3	144–239/240–333/334–428	2×10^{-19}	
SPI4_0/SALTY	Unknown (1512 aa)	*Salmonella typhimurium*	gi|3323600 (1512 aa)	nd	15	[1–19]/20–118/119–215/216–311/312–410/411–512/513–607/608–706/707–799/800–895/896–993/994–1084/1085–1181/1182–1279/1280–1384/[1384–1463]	2×10^{-19}	
Similar but smaller sequences: gi|3323592 (526 aa), gi|3323594 (754 aa), gi|3323595 (172 aa), gi|3323596 (820 aa), gi|3323597 (206 aa), gi|3323599 (92 aa), gi|3323601 (463 aa)								
ESP/ENTFA	Surface protein	*Enterococcus faecalis*	gi|38731187 (1873 aa)	nd	4+9	?–777/778–861/862–945/946–1073/1074–1155/1156–1237/1238–1319/1320–1401/1402–1483/1484–1565/1566–1647/1648–1726/1727–1809	1×10^{-18}	
PSTI/STAST	PstI	*Staphylococcus simulans*	gi|4097699 (173 aa)	nd	2	[1–45]/46–131/[132–173]	2×10^{-15}	
ORF2/SALTY	Proline/threonine–rich protein	*Salmonella typhimurium*	gb|AAD34846.1 (1605 aa)	nd	14	?–272/273–376/377–480/481–564/565–671/672–754/755–840/841–924/925–1035/1036–1139/1140–1242/1243–1346/1347–1448/1449–?	6×10^{-15}	
SLR0364/SYNY3	Hypothetical protein	*Synechocystis sp.*	dbj|BAA10087 (3029 aa)	nd	22	690–795/796–900/901–999/,1000–1097/1098–1196/1197–1295/1296–1393/1394–1492/1493–1591/1592–1689/1690–1788/1789–1887/1888–1985/1986–2084/2085–2183/2184–2281/2282–2380/2381–2479/2480–2577/2578–2676/2677–2775/2776–2900	1×10^{-13}	

SLPO_BACBR	Outer cell wall protein	*Bacillus brevis*	sp|P09333 (1004 aa)	nd	2	?–907/908–1004	5×10^{-13}	
GUXA_CELFI	Exoglucanase A	*Cellulomonas fimi*	sp|P50401 (872 aa)	Fn3	3	479–566/574–664/672–762	2×10^{-7}	CBD GH6
AF011339.1/ACIAD	Unknown	*Acinetobacter sp. ADP1*	gb|AAC27114.1 (918 aa)	nd	4	[1–71]/72–172/173–278/279–366	2×10^{-7}	
GUND_CELFI	Endoglucanase D	*Cellulomonas fimi*	sp|P50400 (747 aa)	Fn3	2	451–540/541–636	2×10^{-6}	GH6
PH0954/PYRHO	4436 aa long hypothetical protein	*Pyrococcus horikoshii*	dbj|BAA30051 (4436 aa)	nd	2	1670–1795/1796–1871	7×10^{-6}	Fn3 PKD
AF1652/ARCFU	Prepro-subtilisin sendai	*Archaeoglobus fulgidus*	gi|2648899 (910 aa)	nd	2	531–610/611–684	3×10^{-5}	Peptidase S8 PKD
SLR0366/SYNY3	Hypothetical protein	*Synechocystis sp.*	dbj|BAA10088 (1742 aa)	nd	9	[1–61]/62–167/168–273/274–379/380–485/486–591/592–720/721–857/858–996	4×10^{-5}	
KIAA0319/HUMAN	KIAA0319	*Homo sapiens*	dbj|BAA2077 (1072 aa)	Fn3/PKD	4	429–523/524–619/620–713/714–810	3×10^{-4}	FN3/PKD
PHD/PSEST	Polyhydroxybutarate depolymerase	*Pseudomonas stutzeri*	dbj|BAA32541 (576 aa)	nd	1	?–442	3×10^{-4}	
BCA_STRAG	C protein alpha–antigen	*Streptococcus agalactiae*	sp|Q02192 (1020 aa)	nd	9	227–306/307–388/389–470/471–552/553–634/635–716/717–798/799–880/881–962	7×10^{-4}	
FAT/SYNY3	Fat protein	*Synechocystis sp.*	dbj|BAA17114 (1965 aa)	nd CA	4 10	348–421/422–525/526–638/639–734/735–836/837–938/939–1040/1041–1142/1143–1244/1245–1346/1347–1448/1449–1550/1551–1652/1653–1754	7×10^{-4}	
CHI1_BACCI	Chitinase A1	*Bacillus circulans*	sp|P20533 (699 aa)	Fn3	2	454–549/550–644	9×10^{-4}	

[a] Searches were limited to this step and not performed until convergence, because HYR domains, particularly rich in "light" amino acids (such as glycine, alanine, serine, and threonine), have biased amino acid composition (also see text), and consequently, PSI-BLAST generates from this step compositionally rooted artefacts, such as that observed for mucins (e.g., MUC1_XENLA, PSI-BLAST E-value 2×10^{-4}). As noted by Altschul and Koonin (1998), such cases can be identified by visual inspection, especially because the hits do not share conserved motifs of the HYR domains. PSI-BLAST E-values for members of the HYR family range from 3×10^{-41} (*S. purpuratus* hyalin) to 2×10^{-17} (W02C12.1/CAEEL). The positions of N- and C-terminal limits are shown with a question mark when not identified precisely.

Proteins composed exclusively of the matching repeated domain are underlined (when the sequence repeats are shifted relative to the structural repeat, N- and C-terminal sequences, which probably participate in a unique structural, unit are shown in parentheses).

Other domains previously identified in the different proteins sequences are also indicated (GH18: Glycosyl Hydrolase family 18; GH6: Glycosyl Hydrolase family 6; CBD: Cellulose-Binding Domain; Peptidase S8: Peptidase Family S8).

[b] nd, not determined.

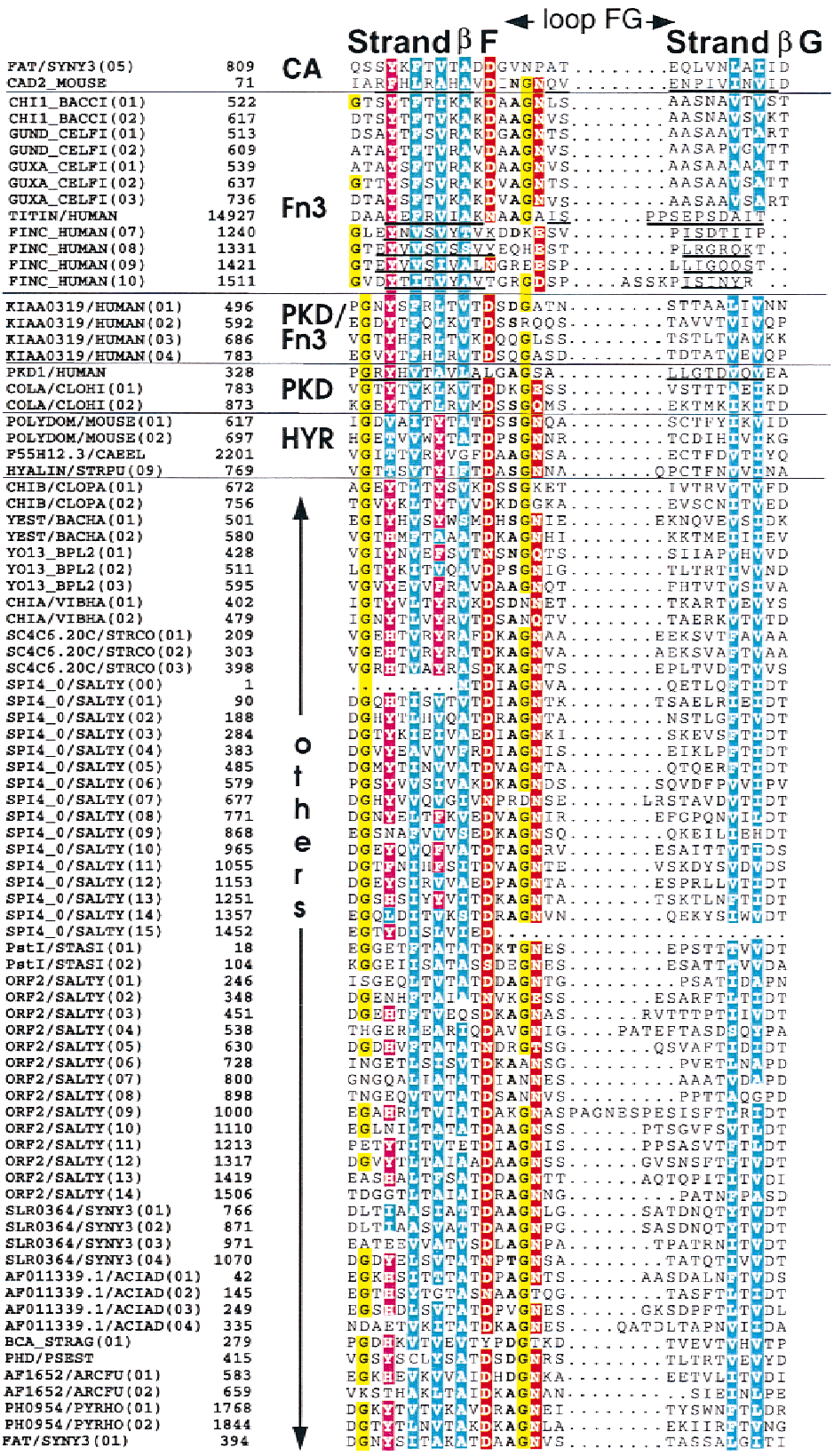

Fig. 4. Alignment of the two C-terminal β-strands of HYR domains and bacterial repeated domains reported in Table 1, relative to those of Fn3 and PKD domains, as well to those of cadherin (CA) domains. The positions of β-strands are underlined when experimentally determined (mouse cadherin (CAD2_MOUSE, PDB 1CNJ); human titin (TITIN/HUMAN, PDB 1BPV); human fibronectin (FINC_HUMAN, PDB 1FNF); human PKD1 (PDK1/HUMAN, PDB 1B4R). Abbreviations are those reported in Table 1, except collagenase from *Clostridium histolyticum* (COLA/CLOHI, pir|I40805).

is justified by the most conserved features of core β-strands (the two sequence-adjacent pairs of strands B/C and E/F, the predicted strands βB, βE, or βF of HYR domains carrying conserved aromatic amino acids), by constraints brought by the probable disulfide bond [linking strand βA and strand βG, like in CD2 (Bodian et al., 1994), and in an Fn3 module of neuroglian (Huber et al., 1994)], and by the conserved motif linking strands βB and βC [similar and conserved motifs are found separating strands βB and βC in chitinase Fn3 domains (Suzuki et al., 1999)].

Bacterial repeated domains as new members within the immunoglobulin-like fold: The bacterial repeated domains, which also should form distinct superfamilies within the immunoglobulin-like fold (Fig. 4; Table 1), are often associated within a same protein sequence with Fn3 (e.g., YesT from *Bacillus halodurans*), CA (e.g., fat protein form *Synechocystis*), and/or PKD domains (e.g., preprosubtilisin sendai from *Archaeoglobus fulgidus*). They are also observed linked to enzymes, which, in other species, possess true Fn3 (e.g., chitinases) or PKD domains (e.g., peptidases). This distribution suggests an evolutionary relationship between these repeated domains and Fn3, PKD, and CA domains, although independent origins for different superfamilies belonging to the immunoglobulin-like fold are generally largely favored (Bork et al., 1994; Shapiro et al., 1995b; Halaby et al., 1999). Some of the bacterial repeated domains highlighted here occur in a number of important proteins such as surface proteins from *streptococci* (Rib and C-alpha protein antigen, ESP protein, R28 protein) known to confer protective immunity (Wästfelt et al., 1996; Shankar et al., 1999; Stalhammar-Carlemalm et al., 1999; Fig. 4). Their characterization relative to the immunoglobulin-like fold thus opens new perspectives for studying their function and for designing efficient vaccines.

In conclusion, while clear sequence similarities are observed between the C-terminal ends of HYR, Fn3, and PKD domains, the HYR family of modules is clearly distinct from classical Fn3 and PKD domains, as the N-terminal sequence hallmarks of these two widespread domains are not conserved in HYR domains, suggesting that the three families might also largely differ in their functions. Although little is known about the function of the HYR module apart from its probable involvement in cellular adhesion, its presence in three proteins from the fully sequenced genome of *Caenorhabditis elegans* (Fig. 2) makes it likely that in the near future, new eukaryotic members will be found and its biological functions elucidated.

Materials and methods: Searches within the nonredundant database (NR) were performed using BLAST2 and PSI-BLAST programs (Altschul et al., 1997) running at the National Center for Biological Information (NCBI, USA). Hidden Markov Model (HMM) searches were carried out using the HMMER package (Eddy, 1998). Databases such as Pfam (Bateman et al., 1999), SMART (Ponting et al., 1999), and Prosite (Hofmann et al., 1999) were also searched for the presence of previously described domains within the protein sequences under investigation.

Bidimensional Hydrophobic Cluster Analysis (HCA) was also used. Guidelines to the use and a review of this method are described elsewhere (Callebaut et al., 1997). HCA combines sequence comparison with secondary structure predictions and is particularly efficient at low levels of sequence identity (typically below 20–25% sequence identity) and in detecting internal repeats (e.g., Callebaut et al., 1999).

Secondary structure predictions were performed using the JPred server (Cuff et al., 1998).

Note added in proof: Since completion of this work, the genome sequence of *Drosophila melanogaster* was published, in which two gene products containing HYR domains were detected (see also http://www.lmcp.jussieu.fr/~callebau/HYR.html):

1. CG7526 (gi|7295215; 1394 aa): 1 HYR domain (aa 1248 to 1329) accompanied by 1 EGF, 14 EGF-CA, and 2 CCP domains.
2. CG9138 (gi|7297206; 3396 aa): 3 HYR domains (aa 1298 to 1384; 1385 to 1468; 2634 to 2716) accompanied by 1 LDLa, 3 CUB, 9 CCP, 2 FA58C, 16 EGF-CA, 2 $\times$ 3 STT2R, 3 EGF, and 1 LamG (laminin G) domains as well as a potential transmembrane segment (aa 3256 to 3276). The domain organization of the first half of the CG9138 gene product is very similar to that of the *C. elegans* F47C12.1 hypothetical protein. The additional information brought by the *D. melanogaster* CG9138 sequence led us to detect a second HYR domain in the F47C12.1 sequence, following the first one described in Figures 2 and 3 (aa 1196 to 1279). The second half of the CG9138 sequence (aa 1842 to 3382) is identical to the recently described SP1070 gene product (gi|7542565; 1551 aa; 100% identity between aa 1 and 1541), whose HYR domain is located between aa 793 and 875.

Acknowledgments: I.C. and J.P.M acknowledge the financial support of the CNRS programs "Physique et Chimie du Vivant" and "Génome."

References

Altschul SF, Koonin EV. 1998. Iterated profile searches with PSI-BLAST—A tool for discovery in protein databases. *Trends Biochem Sci 23*:444–447.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res 25*:3389–3402.

Barlow PN, Steinkasserer A, Norman DG, Kieffer B, Wiles AP, Sim RB, Campbell ID. 1993. Solution structure of a pair of complement modules by nuclear magnetic resonance. *J Mol Biol 232*:268–284.

Baron M, Main AL, Driscoll PC, Mardon HJ, Boyd J, Campbell ID. 1992. ^{1}H NMR assignment and secondary structure of the cell adhesion type III module of fibronectin. *Biochemistry 31*:2068–2073.

Bateman A, Birney E, Durbin R, Eddy SR, Finn RD, Sonnhammer EL. 1999. Pfam 3.1: 1313 multiple alignments and profile HMMs match the majority of proteins. *Nucleic Acids Res 27*:260–262.

Bodian DL, Jones EY, Harlos K, Stuart DI, Davis SJ. 1994. Crystal structure of the extracellular region of the human cell adhesion molecule CD2 at 2.5 Å resolution. *Structure 2*:755–766.

Bork P, Bairoch A. 1995. A proposed nomenclature for the extracellular protein modules of animals. *Trends Biochem Sci 20(3)*:Poster CO2.

Bork P, Doolittle RF. 1992. Proposed acquisition of an animal protein domain by bacteria. *Proc Natl Acad Sci USA 89*:8990–8994.

Bork P, Downing AK, Kieffer B, Campbell ID. 1996. Structure and distribution of modules in extracellular proteins. *Q Rev Biophys 19*:119–167.

Bork P, Holm L, Sander C. 1994. The immunoglobulin fold. Structural classification, sequence patterns and common core. *J Mol Biol 242*:309–320.

Bycroft M, Bateman A, Clarke J, Hamill SJ, Sandford R, Thomas RL, Chotia C. 1999. The structure of a PKD domain from polycystin-1: Implications for polycystic kidney disease. *EMBO J 18*:297–305.

Callebaut I, Courvalin JC, Mornon JP. 1999. The BAH (Bromo-Adjacent Homology) domain: A link between methylation, replication and transcriptional regulation. *FEBS Lett 446*:189–193.

Callebaut I, Labesse G, Durand P, Poupon A, Canard L, Chomilier J, Henrissat B, Mornon JP. 1997. Deciphering protein sequence information through hydrophobic cluster analysis (HCA): Current status and perspectives. *Cell Mol Life Sci 53*:621–645.

Campbell ID, Spitzfaden C. 1994. Building proteins with fibronectin type III modules. *Structure 2*:333–337.

Choudhury D, Thompson A, Stojanoff V, Langermann S, Pinkner J, Hultgren SJ, Knight SD. 1999. X-ray structure of the FimC-FimH chaperone-adhesin complex from uropathogenic *Escherichia coli*. *Science 285*:1061–1066.

Cuff JA, Clamp ME, Siddiqui AS, Finlay M, Barton GJ. 1998. JPred: A consensus secondary structure prediction server. *Bioinformatics 14*:892–893.

Eddy SR. 1998. Profile hidden Markov models. *Bioinformatics 14*:755–763.

Gouet P, Courcelle E, Stuart DI, Metoz F. 1999. ESPript: Analysis of multiple sequence alignments in PostScript. *Bioinformatics 15*:305–308.

Halaby DM, Poupon A, Mornon JP. 1999. The immunoglobulin fold family: Sequence analysis and 3D structure comparisons. *Protein Eng 12*:563–571.

Harpaz Y, Chotia C. 1994. Many of the immunoglobulin superfamily domains in cell adhesion molecules and surface receptors belong to a new structural set which is close to that containing variable domains. *J Mol Biol 238*: 528–539.

Hofmann K, Bucher P, Falquet L, Bairoch A. 1999. The PROSITE database, its status in 1999. *Nucleic Acids Res 27*:215–219.

Huber AH, Wang YM, Bieber AJ, Bjorkman PJ. 1994. Crystal structure of tandem type III fibronectin domains from *Drosophila neuroglian* at 2.0 Å. *Neuron 12*:717–731.

Jones EY. 1996. Three-dimensional structure of cell adhesion molecules. *Curr Opin Cell Biol 8*:602–608.

Jones EY, Harlos K, Bottomley MJ, Robinson RC, Driscoll PC, Edwards RM, Clements JM, Dudgeon TJ, Stuart DI. 1995. Crystal structure of an integrin-binding fragment of vascular cell adhesion molecule-1 at 1.8 Å resolution. *Nature 373*:539–544.

Kraulis PJ. 1991. Molscript—A program to produce both detailed and schematic plots of protein structures. *J Appl Crystallogr 12*:283–291.

Leahy DJ, Hendrickson WA, Aukhil I, Erickson HP. 1992. Structure of a fibronectin type III domain from tenascin phased by MAD analysis of selenomethyionyl protein. *Science 258*:987–991.

Little E, Bork P, Doolittle RF. 1994. Tracing the spread of fibronectin type III domains in bacterial glycohydrolases. *J Mol Evol 39*:631–643.

Meindl A, Carvalho MRS, Herrmann K, Lorenz B, Achatz H, Lorenz B, Apfelstedt-Sylla E, Wittwer B, Ross M, Meitinger T. 1995. A gene (SRPX) encoding a sushi-repeat-containing protein is deleted in patients with X-linked retinitis pigmentosa. *Hum Mol Genet 4*:2339–2346.

Overduin M, Harvey TS, Bagby S, Tong KI, Yau P, Takeichi M, Ikura M. 1995. Solution structure of the epithelial cadherin domain responsible for selective cell adhesion. *Science 267*:386–389.

Ponting CP, Schultz J, Milpetz F, Bork P. 1999. SMART: Identification and annotation of domains from signalling and extracellular protein sequences. *Nucleic Acids Res 27*:229–232.

Reid KBM, Day AJ. 1989. Structure–function relationships of the complement components. *Immunol Today 6*:177–180.

Sauer FG, Futterer K, Pinkner JS, Dodson KW, Hultgren SJ, Waksman G. 1999. Structural basis of chaperone function and pilus biogenesis. *Science 285*:1058–1061.

Shankar V, Baghdayan AS, Huycke MM, Lindahl G, Gilmore MS. 1999. Infection-derived *Enterococcus faecalis* strains are enriched in esp, a gene encoding a novel surface protein. *Infect Immun 67*:193–200.

Shapiro L, Fannon AM, Kwong PD, Thomson A, Lehmann MS, Grubel G, Legrand JF, Alsnielsen J, Colman DR, Hendrickson WA. 1995a. Structural basis of cell cell adhesion by cadherins. *Nature 374*:306–307.

Shapiro L, Kwong PD, Fannon AM, Colman DR, Hendrickson WA. 1995b. Considerations on the folding topology and evolutionary origin of cadherin domains. *Proc Natl Acad Sci USA 92*:6793–6797.

Stalhammar-Carlemalm M, Areschoug T, Larsson C, Lindahl G. 1999. The R28 protein of Streptococcus pyogenes is related to several group B streptococcal surface proteins, confers protective immunity and promotes binding to human epithelial cells. *Mol Microbiol 33*:208–219.

Suzuki K, Taiyoji M, Sugawara N, Nikaidou N, Henrissat B, Watanabe T. 1999. The third chitinase gene (chiC) of Serratia marcescens 2170 and the relationship of its product to other bacterial chitinases. *Biochem J 343*:587–596.

Wästfelt M, Stälhammar-Carlemalm M, Delisse AM, Cabezon T, Lindahl G. 1996. Identification of a family of streptococcal surface proteins with extremely repetitive structure. *J Biol Chem 271*:18892–18897.

Wessel G, Berg L, Adelson DL, Cannon G, McClay DR. 1998. A molecular analysis of hyalin, a substrate for cell adhesion in the hyaline layer of the sea urchin embryo. *Dev Biol 193*:115–126.